Supplementary Material

Linear effects models of signaling pathways from combinatorial perturbation data

Ewa Szczurek[1], Niko Beerenwinkel[2,*]
[1] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland,
[2] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland; SIB Swiss Institute of Bioinformatics
[*] E-mail: niko.beerenwinkel@bsse.ethz.ch

Proof. **(Lemma 1)** Consider a given model graph $\mathcal{G}$ uniquely defined by its binary adjacency matrix. Recall that the adjacency matrix has entry 1 in row i and column j if and only if there is a directed edge from i to j in $\mathcal{G}$. Thus, each entry (i, j) in the adjacency matrix corresponds to the family relation of the ordered pair of nodes (i, j). Value 1 in entry (i, j) is equivalent to the fact that i is a parent of j, while entry 0 is equivalent to the fact that i and j are cousins. □

Proof. **(Lemma 2)** Assume the contrary. We will show by example that this assumption does not hold. Let A and B be two LEMs on $G = \{1, 2\}$, one of which has the nodes *1* and *2* in a parent-child relation, and another has *1* and *2* as cousins (illustrated by example in Figure 2A-D in the main text). Consider the two single node perturbation experiments on these genes, $\{1\}$ and $\{2\}$, where the first experiment targets gene *1* and second gene *2*. The states matrix for model A with these experiments is

$$S^A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix},$$

while the states matrix for B is

$$S^B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Consider the parameters $\beta^B = [\beta_1, \beta_2]^T$ and $c^B = c$ for model B. Model A with parameters $\beta^A = [\beta_1, \beta_2 - \beta_1]^T$ and the same precision parameter $c^A = c$ will have equal likelihood. Indeed, with those parameter values $S^A\beta^A = S^B\beta^B = [\beta_1, \beta_2]^T$. □

Proof. **(Lemma 3)** Assume the contrary. We will show by example that this assumption does not hold. As illustrated in Figure 2E-G in the main text, we consider two LEMs on $G = \{1, 2\}$, one of which has nodes *1* and *2* in a parent-child relation, and another has *1* and *2* as cousins. Consider the only possible double perturbation experiment, $\{1, 2\}$, where the two nodes are perturbed at the same time. In this case, for both models, the states matrix will be the same and will consist of a single row, $S = [1, 1]$. Thus for any contribution vector $[\beta_1, \beta_2]^T$ and precision c, we will have $S^A\beta = S^B\beta = S\beta = \beta_1 + \beta_2$, and for $c^A = c^B = c$ the likelihood of the two different models will be the same. □

Proof. **(Theorem 1)** By Lemma (1), to uniquely identify a LEM on G, for each pair of nodes in G their family relation needs to be specified. For any given pair of nodes *1, 2*, there are two general ways in which an alternative model can be wrong regarding their relation:

error (1) *1* and *2* are in a parent-child relation, but are taken for cousins, or vice versa,

error (2) *2* is a parent of *1*, but is taken for a child of *1*, or vice versa.

We show that with both perturbations of all single and all pairs of nodes in G, i.e., also both single and pairwise perturbations of *1* and *2*, none of these two errors is possible.

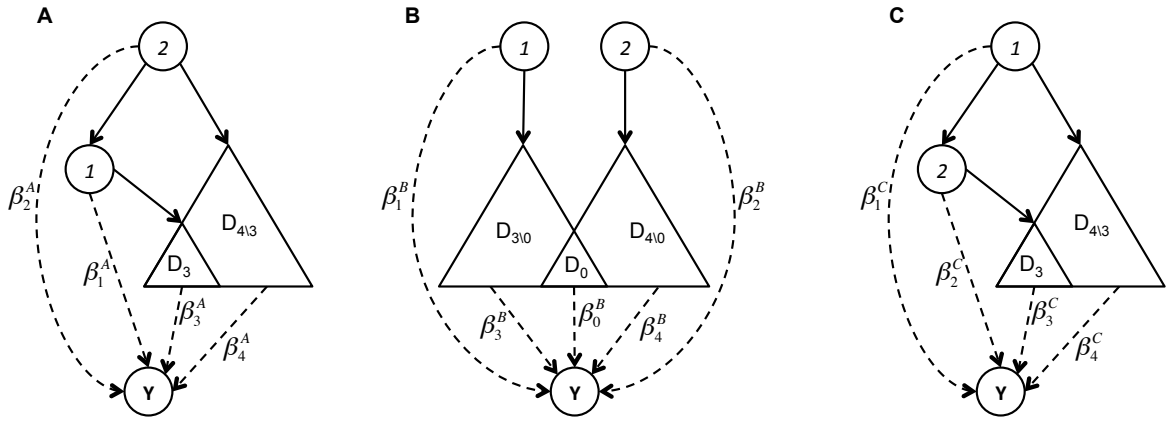

Figure 1. Alternative LEMs with regard to relation between *1* and *2*.

We will consider the three alternative types of LEMs, A, B and C, presented in Figure 1, in which genes *1* and *2* can be related to each other in different ways. In model A, *2* is a parent of *1*, and their corresponding contribution parameters are β_1^A, and β_2^A. This model is presented in Figure 1A. Model A corresponds to a general scenario, where *1* and *2* can be a part of a larger graph, with a set of nodes D_3 which are descendants of node *1* and a set of nodes $D_{4\setminus 3}$ which are descendants of *2* but do not belong to D_3 (by transitive closure D_3 is a subset of all descendants of *2*). Whether nodes *1* or *2* have parents is irrelevant for our discussion. Denote the sum of contributions of descendants of node *1* as $\beta_3^A = \sum_i [i \in D_3]\beta_i$, where $[\cdot]$ represents the indicator function. The sum of contributions of descendants of *2* in $D_{4\setminus 3}$ is denoted β_4^A. Recall that $(S_e)^T$ denotes the row vector of the states matrix for experiment e. From Fact 1, the entries of $S^A\beta^A$ for model A for the set of single and double perturbation experiments on *1* and *2* are given by

$$(S^A_{\{1\}})^T\beta^A = \beta_1^A + \beta_3^A \quad (1)$$

$$(S^A_{\{2\}})^T\beta^A = \beta_1^A + \beta_2^A + \beta_3^A + \beta_4^A \quad (2)$$

$$(S^A_{\{1,2\}})^T\beta^A = \beta_1^A + \beta_2^A + \beta_3^A + \beta_4^A. \quad (3)$$

In another LEM, denoted B and presented in Figure 1B, *1* and *2* are cousins. In this model, node *1* has contribution β_1^B, while *2* has contribution β_2^B. Their sets of descendants may overlap. D_0 denotes the set of overlapping descendants of *1* and *2* in B, while $D_{3\setminus 0} = D_3 \setminus D_0$ and $D_{4\setminus 0} = D_4 \setminus D_0$. The sums of corresponding descendant contributions are denoted β_0^B, β_3^B, and β_4^B. The entries of $S^B\beta^B$ for model B for the set of single and double perturbation experiments are given by

$$(S^B_{\{1\}})^T\beta^B = \beta_1^B + \beta_3^B + \beta_0^B, \quad (4)$$

$$(S^B_{\{2\}})^T\beta^B = \beta_2^B + \beta_4^B + \beta_0^B, \quad (5)$$

$$(S^B_{\{1,2\}})^T\beta^B = \beta_1^B + \beta_2^B + \beta_3^B + \beta_4^B + \beta_0^B. \quad (6)$$

$$(7)$$

Finally, as the third possibility, in LEM C, presented in Figure 1C, *1* is a parent of *2*, and their contributions are denoted β_1^C, and β_2^C, respectively. In this model, the set of descendants of *2* is denoted D_3, and the set of descendants of *1*, not belonging to D_3, is denoted $D_{4\setminus 3}$. The sum of contributions of descendants of *2* in D_3 is denoted β_3^C, while the sum of contributions of descendants of *1* contained in $D_{4\setminus 3}$ is denoted β_4^C. For this model,

the expected effects with single and double node perturbations of *1* and *2* are the following

$$(S^C_{\{1\}})^T\beta^C = \beta^C_1 + \beta^C_2 + \beta^C_3 + \beta^C_4 \quad (8)$$
$$(S^C_{\{2\}})^T\beta^C = \beta^C_2 + \beta^C_3 \quad (9)$$
$$(S^C_{\{1,2\}})^T\beta^C = \beta^C_1 + \beta^C_2 + \beta^C_3 + \beta^C_4. \quad (10)$$

We will now exclude that error (1) occurs due to equal likelihoods. Assume first that LEM A and the alternative LEM B (Figure 1) have equal likelihood. To obtain equal likelihood, $S^A\beta^A = S^B\beta^B$ and $c^A = c^B$ must be satisfied. Assume $c^A = c^B$. Consider two experiments: single perturbation of *2*, and double perturbation of *1* and *2*. For the two models A and B to have equal likelihood with this set of experiments, we should have

$$(S^A_{\{2\}})^T\beta^A = \beta^A_1 + \beta^A_2 + \beta^A_3 + \beta^A_4 = (S^B_{\{2\}})^T\beta^B = \beta^B_2 + \beta^B_4 + \beta^B_0$$
$$(S^A_{\{1,2\}})^T\beta^A = \beta^A_1 + \beta^A_2 + \beta^A_3 + \beta^A_4 = (S^A_{\{1,2\}})^T\beta^B = \beta^B_1 + \beta^B_2 + \beta^B_3 + \beta^B_4 + \beta^B_0$$

By subtracting the first from the second equation, this amounts to

$$\beta^B_1 + \beta^B_3 = 0,$$

which contradicts the model assumption that $\beta^B > 0$.

Note that model C in Figure 1 is symmetric to A, and contradiction to the assumption that models B and C could have equal likelihoods would follow the same lines of argumentation.

We will now exclude that error (2) occurs due to equal likelihoods. Assume first that LEM A, where *2* is a parent of *1* and LEM C, where *1* is a parent of *2* have equal likelihood, and $c^A = c^C$. Consider two single node perturbations, of *1* and of *2*. For the two models A and C to have equal likelihoods we should have

$$(S^A_{\{1\}})^T\beta^A = \beta^A_1 + \beta^A_3 = (S^C_{\{1\}})^T\beta^A\beta^C_1 + \beta^C_2 + \beta^C_3 + \beta^C_4$$
$$(S^A_{\{2\}})^T\beta^A = \beta^A_1 + \beta^A_2 + \beta^A_3 + \beta^A_4 = (S^C_{\{2\}})^T\beta^C\beta^C_2 + \beta^C_3.$$

This is equivalent to

$$\beta^A_2 + \beta^A_4 = -(\beta^C_1 + \beta^C_4),$$

which contradicts the assumption that $\beta^A > 0$ and $\beta^C > 0$.

In summary, we have shown that for any pair of nodes *1, 2*, and with both single and double experiments on those nodes, their family relation can uniquely be estimated from the data, in the sense that there exists no alternative model with equal likelihood where this relation is different. Thus, by Lemma (1), the LEM is identifiable. □